PREDICTION OF RATING POINT USING MACHINE LEARNING ALGORITHMS

N.Kanimozhi[1]
Assistant Professor
Department of Computer Technology-UG
Kongu Engineering College
Erode, TamilNadu
kanimozhi6465@gmail.com

M.N.Kavitha[2]
Assistant Professor
Department of Computer Technology-UG
Kongu Engineering College
Erode, TamilNadu
kavithafeb1@gmail.com

S.S.Saranya[3]
Assistant Professor
Department of Computer Technology-UG
Kongu Engineering College
Erode, TamilNadu
saranya.soundarajan@gmail.com

Aravinth S[4]
Student, Department of Computer Technology-UG
Kongu Engineering College
Erode, TamilNadu
aravinths.19bcr@kongu.edu

Kavin Prakash M[5]
Student, Department of Computer Technology-UG
Kongu Engineering College
Erode, TamilNadu
kavinprakash.mkp@gmail.com

Naren D.K[6]
Student, Department of Computer Technology-UG
Kongu Engineering College
Erode, TamilNadu
narendk.19bcr@kongu.edu

Abstract— **Film Industry is not only an industry or a centre of entertainment, rather now it's became a centre of global business. All over the world are now excited about a movie's box office success rate, popularity etc. A huge dataset is available across online platforms about the movie's success rate. We have used Hollywood movie list from Wikipedia and their rating from the IMDb movie rating website to create our data set. IMDb (Internet Movie Database) is an online database of information related to films, television series, and streaming content online – including cast, production crew, personal biographies, plot summaries, ratings, fan, and critical reviews.**

It's hard to predict the movie success rate before a movie is released. Many machine learning techniques can be used to predict the success rate. Here, we are using regression algorithms to predict success rate. An efficient model is developed using Random Forest, Decision Tree, Support vector machine, and XGBoost to predict movie's success rate.

Keywords- imdb score, movie ratings, regression, decision tress, random forest, xgboost, svm..

I. INTRODUCTION

Nowadays movies are not only a source of entertainment and recreation, but it is also the main source of global marketing. Movies create a new craziness among people, especially among the younger generation. The success of the movie is not only dependent on directors and box office collection but also on the reach of the movie to general people. People talk about reviews of movies on social media. Therefore, people analyze movies on social media other than that they analyze some other scopes like the director's previous movie success and the actor's popularity. Analysis may be different among different countries. People from different regions in the world react uniquely. People's reviews on movies are now available on the internet. There are platforms like IMDb, which help people to find reviews on the success of movies.

II. RELATED WORK

Ajinkya Ambadkar, Rahul Jojare , Rushal Wankhade proposed work on Prediction and Sentimental Analysis of TRP rating Using Artificial Intelligence Approach [1] (2020).The number of TV shows that are produced each year is increasing at an alarming rate, which means that the success rate of these shows is very important for the industry. This is why the prediction of such shows is very important. We will be using the data collected by Kaggle.com, which is a database that contains ratings of over a thousand Netflix shows.

Haroon Malik, M.Shakshuki, and Ansar-Yasar proposed work on Approximating viewership of streaming T.V Programs using social media sentiment analysis [2] (2021). Using social media data, the research teams developed a technique to estimate the number of viewers of tv programs on the Internet. Their method was able to achieve an accuracy of 85%.

Muhammad Sudais, Mohammad Hasan Khan, and Abdul Jabbar Tabani researched on Performance Prediction for IMDB Movies [3] (2022). Filmmakers are very concerned about the box office performance of their movies. They spend a lot of money on their movies and are hoping that the audiences will give them a good review. They also want to receive healthy nominations and award wins.

Ruihan Hu proposed a work on TV Series Ratings Analysis and Prediction Based on Decision Tree [4] (2019). Through various experiments, the paper compared the results of the two models and came up with a decision tree algorithm that is known as ID3. The results show a good effect of the prediction model constructed in this paper.

Rajeswari, Prasad, and Kiran presented a paper An Advanced Neighbourhood approach of recommending movies on Netflix data by the combination of KNN and XGBoost [5] (2020). The researchers then used the recommendation, nearest neighbor, and XGBoost to analyze the ratings of various movies on Netflix. They discovered that almost all of them were 99% of ratings not filled due to the low number of people watching them. By using the neighbor model as the factor in the developed models KNN and XGBoost, the work was gradually constructed with various components of the model by 8 formulations. Using the KNN-XGBoost model can be applied to predict rating.

The precision of the proposed system is 82 % while using svd most accurate will be found as the future enhancement.

Author name Saura Sambit Acharya, Ashvin Gupta, and Prabu Shankar K.C [6] (2019) proposed work on TV Show Popularity Analysis using social media, Data Mining. In this paper, the author used the K-nearest neighbors algorithm (KNN), Support Vector Clustering (SVM). This Algorithm is used to classify the Popularity factors of various types of movies like thrillers and drama. That helps in the movie's success prediction. The datasets used in this project to train machine learning models are obtained from the IMDB site. The object of this paper is to find the popularity of the tv show using social media and data mining. Their future enhancement is to obtain input from the audience which can be added to the dataset to improve the result.

Sidhu, Attwal, and R.Gaurav presented To Evaluate & Predict the Television Serials‟ TRP [7] (2019). Day by day the number of TV shows is increasing. viewers' opinions are examined first. After examining the viewer's opinion that six factors have been constructed ie .. a random tree, This random tree is a part of the classification. an iterative learning revelation process like Data cleaning, Data coordination, Data selection, Data change, and Data mining was made. Techniques used Classification: a technique that assumes items in a collection to target categories or classes and Prediction: to identify data points purely on with help of another related data value. These 6 major factors were considered so more factors can be added.

Hirofumi Nonaka, Asahi Hentona, and Hirochika Yamashiro presented about Measuring the influence of the mere exposure effect of TV commercial adverts on purchase behavior based on machine learning prediction models [8] (2019). Due to the rise of new information technologies and the increasing popularity of online advertisements, many studies have been conducted on the effects of online and television advertisements on purchase behavior. However, they applied machine learning algorithms SVM and XGBoost and they found that the effects of ad exposure time on the purchase behavior were not significant.

L.M.Tadesse, Hongfei proposed work on Personality Predictions Based on User Behavior on the Facebook Social Media Platform [9] (2019). Due to the rise of social networking sites, various approaches have been developed to analyze the personalities of users. This paper aims to analyze the various features of social networks and their relationship with their personality traits. Through four machine learning models, the researchers were able to perform a statistical analysis of the data.

Dipak Gaikar, Riddhi Solanki, and Harshada Shinde proposed work on IRJET- Movie Success Prediction Using Popularity Factor from social media [10] (2019). The film industry has been experiencing immense growth over the past couple of years. Due to the immense amount of business that the industry has been able to generate, it has been difficult for people to predict the success of various movies. This project aims to help the users make informed decisions regarding the tickets they buy by analyzing data from various sources.

III. EXISTING SYSTEM

The taken dataset consists of some missing values, incorrect values hence to overcome this type of problem is to clean and refine the data. K-NN is used for the classification of data, it classifies data by getting majority votes from its neighbors. Prediction is done for various. The predicted rating is compared with the IMDB rating to find the accuracy percent. The bar graph is plotted with an accurate result. The prediction of movie ratings helps in forecasting the success of the future upcoming movies in the earliest.

IV. EXISTING SYSTEM DRAWBACKS

1.Minimum time needed for various processing so that it makes the users to interact more effectively.

2.Better service and good updates are given when compared with the previously existing work.

3.In previous existing work TRP were found without classifying the movie.

V. PROPOSED SYSTEM

The input from the dataset is initially pre-processed using machine learning techniques such as filter and wrapper in order to reduce undesirable and multiple data values in the suggested work. This machine learning technique shrink the dimensional of the input data by cleaning the input vales. then, the data is divided into testing data and training data. The training data is again divided based on the TV show category .Next, the TRP for each channel is obtained, by calculating the number of users level, user ratingand user rating size. The attributes taken here for calculation is user level, user rating and user rating size. The model was created with the Java programming language

VI. ADVANTAGES

1)The selected regression algorithm plays a vital advantage.

2)The accuracy produced increases when compare to other models in machine learning techniques.

3)huge amount data can also be calculated.

VII. MODULE

A. Dataset

The dataset for this project was obtained from kaggle.com, a popular platform for machine learning professionals. Users of Kaggle can be able to search for a dataset and models which are published. Our dataset is based on IMDb movie rating reviews and is a numeric datatype. It has 5000 movie reviews divided into two categories: favourable and negative.

B. Random forest regression:

Let's make the algorithm simple and clear

1. You have your original dataset df, you want to have K number of Decision Trees in our ensemble. You also have a number N – you will build a Tree until each node has fewer or equal to N samples. you have a number F - the number of features that will be chosen at random in each node of the Decision Tree.

2. First, the original dataset df is taken by the Random Forest and divides it into K number of subsets. "Out-of-bag" samples are those that do not appear in any subset.

3. K number of trees are constructed from a single subset. Further, each tree is constructed until there are fewer or equal to N samples in each node. Furthermore, F features are chosen at random in each node. One is used to split the node.

4. The ensemble of K trained models produces the final result for the Regression task by averaging the predictions of the individual trees.

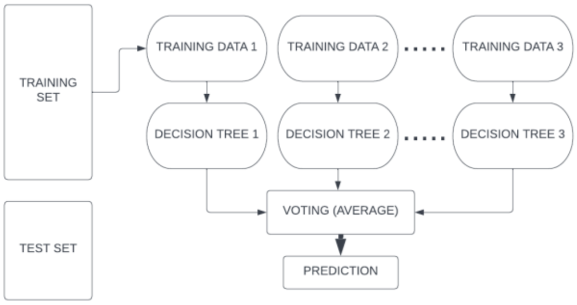

Fig.1 RF REGRESSION

C. Decision tree:

A Decision tree is structured like a flowchart. A decision tree is created parallelly when the dataset is divided into smaller and smaller subsets at the same time. In which, there are three nodes i.e. root node, interior node, and leaf node. Test future represented on each interior node, class label represented on each leaf node and branches lead to those class labels.

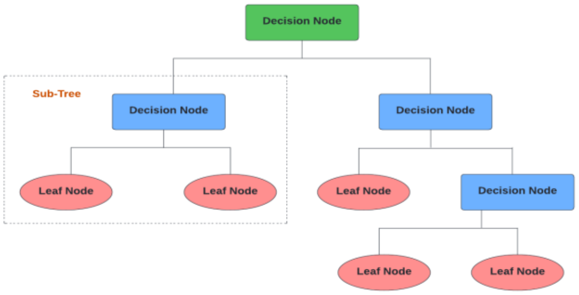

Fig.2 DECISION TREE

Important Terminology:

- Splitting: this purpose is to divide the node into more than one sub-node
- Pruning: removal of sub-node from decision node

D. Support vector machine:

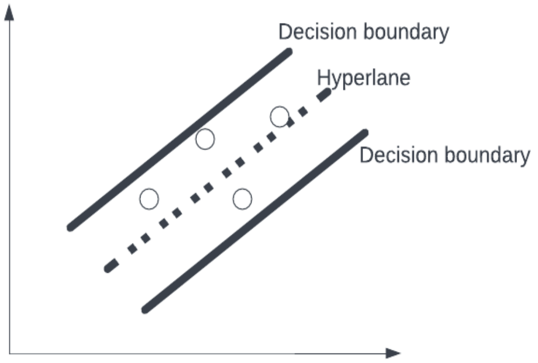

FIG.3 HYPERLANE

With a few small differences, the Support Vector Regression (SVR) does use the same aspects as the SVM for classification. For beginners, because the output is a true figure, considering the future this same information available, which has an unlimited number of possibilities, becomes extremely difficult.

In the regression problems, a margin of compassion is provided as an unbiased estimator to the SVM which the problem has already requested. Besides, there is a much more difficult reason to consider: the method is more complicated. The basic idea, however, remains the same, reducing the error by customizing the hyperplane to maximize the margin while taking into account the constraints.

E. xgboost:

XGBoost can be called Extreme Gradient Boosting. It is more accurate to find the best tree model. It is an efficient gradient boosting implementation to use for regression predictive modeling.

Boosting is the technique of ensembles in which earlier model errors are corrected in new models. Its model is designed to define the connection here between predictor and outcome factors using all available data, which is then generalized to the test data.

F. Prediction:

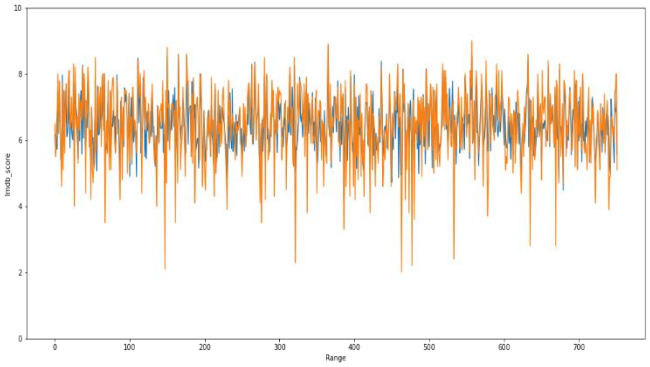

Fig.4 RANDOM FOREST PREDICTION

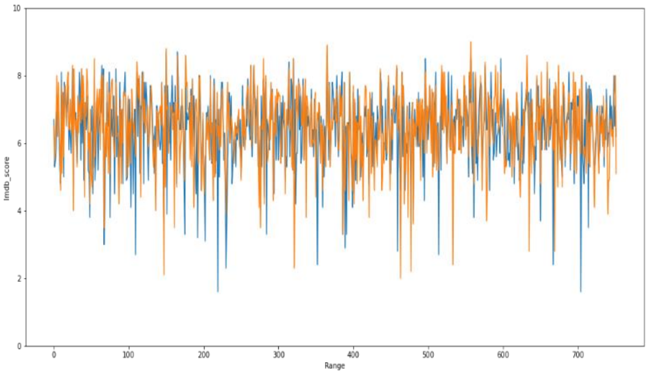

Fig.5 DECISION TREE PREDICTION

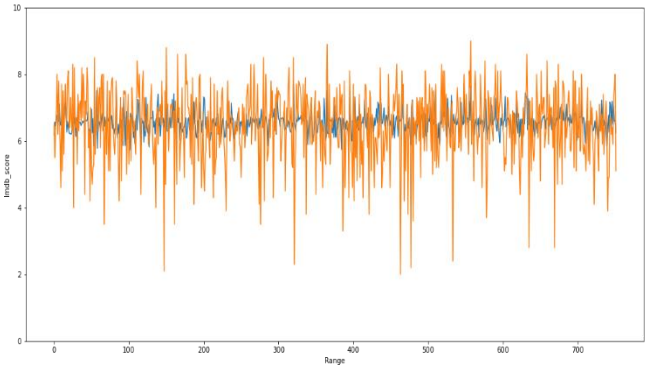

Fig.6 SUPPORT VECTOR MACHINE PREDICTION

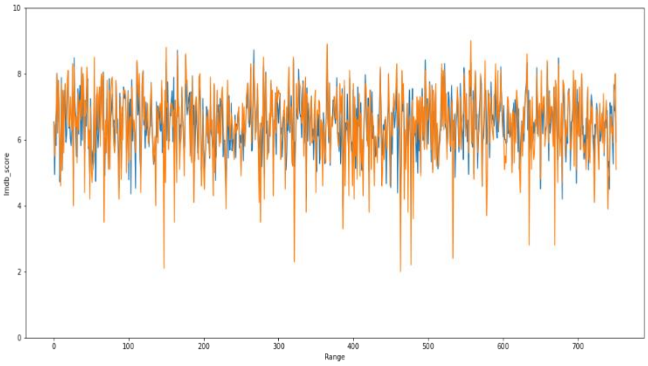

Fig.7 XGBOOST PREDICTION

VIII. CONCLUSION

The film industry has been experiencing immense growth over the decade. Due to the immense amount of business that the industry has been able to predict, it has been difficult for people to predict the success of various movies. This project aims to help the users make informed decisions regarding the tickets they buy by analyzing data from various sources.

Various algorithms such as a random forest, a decision tree, and a support vector machine have also been regarded as highly accurate. XGBoost is a sentiment analysis tool that can be used to analyze and predict the success of various movies. It has been shown that it can achieve an accuracy of 91%.

TABLE 1 ACCURACY

MODEL	**ACCURACY**
XGBOOST	91.187337
RANDOM FOREST	90.639434
DECISION TREE	87.666490
SUPPORT VECTOR MACHINE	85.771786

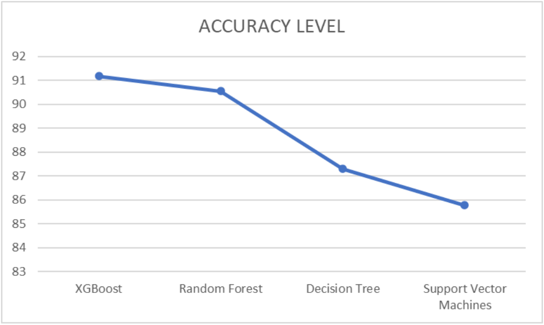

Fig.5 ACCURACY LEVEL

IX. FUTURE ENHANCEMENT

The number of online users were increased day by day. The reviews give the exact standard of the movies. Online review comments are one of the key factors to express feelings about the movie. From the opinion of the reviewers, IMDb will release the rating of the movie. It is a long process to classify the result of the comment. so that instead of getting typing comments from the users, get updated to an emoji review system. For every particular emoji, there will be a specific rating, so that calculation for the IMDb is much better and easy to classify.

REFERENCES

[1] Ajinkya Ambadkar, Rahul Jojare , Rushal Wankhade “Prediction and Sentimental Analysis of TRP rating Using Artificial Intelligence Approach”,Volume 9, Issue 4, April, 2020.

[2] Haroon Malik, M.Shakshuki, Ansar-Yasar “Approximating viewership of streaming T.V program using social media sentiment analysis”,The 12th International Conference on Emerging Ubiquitous Systems and Pervasive Networks(EUSPN), vol 198, november, 2021.

[3] Muhammad Sudais, Mohammad Hasan Khan, and Abdul Jabbar Tabani “Performance Prediction for IMDB Movies “, DOI:10.21203/rs.3.rs-1243202/v1, Janunary,2022

[4] Ruihan Hu “TV Series Ratings Analysis and Prediction Based on Decision Tree “, 6th International Conference on Robotics and Artificial Intelligence, 2019.

[5] Rajeswari, Prasad, and Kiran “An Advanced Neighbourhood Approach of recommending movies on Netflix data by the combination of KNN and XGBoost”,journal of critical reviews, vol 7, Issue 12, 2020.

[6] SauraSambitAcharya,Ashvin Gupta. “TV Show Popularity Analysis using Social Media,Data Mining” International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol 8, May 2019

[7] Sidhu, Attwal and R. Gaurav “To Evaluate & Predict the Television Serials TRP”, International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 6, Issue 9, August ,2019.

[8] Hirofumi Nonaka, Asahi Hentona, Hirochika Yamashiro “ Measuring the influence of the mere exposure effect of TV commercial adverts on purchase behavior based on machine learning prediction models”, 2019

[9] L.M.Tadesse, Hongfei “Personality Predictions Based on User Behavior on the Facebook Social Media Platform”, IEEE Access,volume 6, 2018.

[10] Dipak Gaikar, Riddhi Solanki, Harshada Shinde “IRJET- Movie Success Prediction Using Popularity Factor from Social Media”, International Research Journal of Engineering and Technology (IRJET), vol 6, April, 2019.

[11] Tejaswi Kadam, Gaurav Saraf. “Tv Show Popularity using Sentiment Analysis”, International Research Journal of Engineering and Technology (IRJET), Vol 4, November, 2017.

[12] N.Krishnamoorthy,K.S Ramya. “TV Show Popularity and Performance Predication Using CNN Algorithm”, Journal of Adv Research in Dynamical & Control Systems, Vol 12, 2020.

[13] D.Anand,A.V.Satyavani. “Analysis And Prediction Of Television Show Popularity Rating Using Incremental K-Mean Algorithm”, International journal of mechanical engineering and technology (IJMET) vol 9, January, 2018.

[14] Bane and Sheetlani. “Success of Bollywood Movies Using Machine Techniques” Mukt Shabd Journal, Vol IX, October, 2020.

[15] Vijayakumar, T. "Comparative study of capsule neural network in various applications." Journal of Artificial Intelligence 1, no. 01 (2019): 19-27

[16] Anand, C. "Comparison of Stock Price Prediction Models using Pre-trained Neural Networks." Journal of Ubiquitous Computing and Communication Technologies (UCCT) 3, no. 02 (2021): 122-134

[17] Andi, Hari Krishnan. "An Accurate Bitcoin Price Prediction using logistic regression with LSTM Machine Learning model." Journal of Soft Computing Paradigm 3, no. 3 (2021): 205-217.